A Canonical Form for Flexible Multiword Expressions

Jan Odijk, Martin Kroon
Utrecht University
Utrecht, the Netherlands
j.odijk@uu.nl, m.s.kroon@uu.nl

Abstract

This paper proposes a canonical form for Multiword Expressions (MWEs), in particular for the Dutch language. The canonical form can be enriched with all kinds of annotations that can be used to describe the properties of the MWE and its components. It also introduces the DUCAME (DUtch CAnonical Multiword Expressions) lexical resource with more than 11k MWEs in canonical form. DUCAME is used in MWE-Finder to automatically generate queries for searching for flexible MWEs in large text corpora.

Keywords: multiword expressions, Dutch, canonical form, design principles, automatic query generation, searching for multiword expressions

1. Introduction

This paper proposes design principles and guidelines for a canonical form for (flexible) Multiword Expressions (MWEs) and introduces a lexical resource, DUCAME, which contains more than 11k Dutch multiword expressions in the canonical form as defined in this paper. Most design principles and guidelines are applicable to any language, though some aspects are language-specific and have been specified here for Dutch.

A canonical form for a MWE is a unique representation for a set of variants of this MWE that differ only in grammatical properties. A canonical form for MWEs is necessary because many MWEs are flexible, i.e., their component words can occur in different forms, in different orders, or need not always be adjacent. Until now, no well-defined canonical form for MWEs exists for Dutch (or, to the best of our knowledge, for any other language), though there may be for certain subclasses of MWEs. Single words already have a canonical form, called *lemma*, which is used as a headword in traditional dictionaries. Traditional dictionaries do not use a canonical form but an example to illustrate MWEs. It is already difficult for humans to determine the properties of an MWE on the basis of an example, but it is completely impossible for software. And apart from that, we want to obtain a formal description of the language, and also of MWEs, so the lexical description of MWEs must be as precise as possible.

The canonical form can be enriched with all kinds of annotations, as will be described in detail in this paper. These annotations are necessary to describe the properties of the MWE and its components. The canonical form of the MWE can also serve as a 'headword' for lexical entries in a MWE lexicon, and additional properties of a MWE can be described in its lexical entry.

The canonical form, with its annotations, enables the automatic generation of queries to search for MWEs in large text corpora.[1] This has been implemented in MWE-Finder (Odijk et al., 2023), integrated in GrETEL Version 5 (Odijk et al., to appear 2024),[2] which includes DUCAME.

The structure of this paper is as follows. Section 2 introduces the notion of Multiword Expression, Section 3 discusses related work, Section 4 introduces the DUCAME resource, and Section 5 defines the canonical form for MWEs. Section 6 describes how to deal with properties that are not specific to the MWE, and Section 7 concludes.

2. Multiword expressions

A multiword expression (MWE) is a word combination with linguistic properties that cannot be predicted from the properties of the individual words or the way they have been combined by the rules of grammar (Odijk, 2013b).[3] We will call the individual words that make up a MWE the MWE's *components*. A word combination can, for example, have an unpredictable meaning (*de boeken neerleggen*, lit. 'to put down the books', meaning 'to declare oneself bankrupt'), an unpredictabe form (e.g. *ter plaatse* 'on location', with idiosyncratic use of *ter* and *e*-suffix on the noun), or it can have only limited usage (e.g. *met vriendelijke groet* 'kind regards', used as the closing of a letter).

Words of a MWE need not always be fixed. This can be illustrated with the Dutch MWE *de boeken neerleggen* 'to declare oneself bankrupt'. The verb *neerleggen* in (1) can occur in all of its inflectional

[1] These queries can identify potential occurrences of MWEs. They cannot distinguish an idiomatic reading from a literal reading.

[2] https://gretel5.hum.uu.nl/

[3] For a similar but slightly different definition see (Sag et al., 2001).

LREC-COLING 2024, pages 91–101

variants (e.g., past participle in (1a), infinitive in (1b), and past tense singular in (1c) and (1d)) , and with the separable particle *neer* attached to it (1a, 1b) or separated (1c, 1d). MWEs do not necessarily consist of words that are adjacent, and the words making up a MWE need not always occur in the same order. This expression allows an order with contiguous elements (as in (1a)), but it also allows other words to intervene between its components (as in (1b)), as well as permutations of its component words (as in (1c)), and combinations of permutations and intervention by other words that are not components of the MWE (as in (1d)):[4]

(1) a. Saab heeft gisteren **de boeken**
Saab has yesterday the books
neergelegd.
down.laid
'Saab declared itself bankrupt yesterday.'

b. Ik dacht dat Saab gisteren **de**
I thought that Saab yesterday the
boeken wilde **neerleggen**.
books wanted down.lay
'I thought Saab wanted to declare itself bankrupt yesterday.'

c. Saab **legde de boeken neer**.
Saab laid the books down
'Saab declared itself bankrupt.'

d. Saab **legde** gisteren **de boeken**
Saab laid yesterday the books
neer.
down
'Saab declared itself bankrupt yesterday.'

In addition, certain MWEs allow for (and require) controlled variation in lexical item choice, e.g. in expressions containing bound anaphora such as *zijn geduld verliezen* 'to lose one's temper', where the possessive pronoun varies depending on the subject (cf. *Ik verloor mijn/*jouw geduld; jij verloor *mijn/jouw geduld*, etc.), exactly as the English expression *to lose one's temper*. Of course, not every MWE allows all of these options, and not all permutations of the components of a MWE are well-formed (e.g. one cannot have **Saab heeft neergelegd boeken de.* lit. 'Saab has downlaid books the.').

This flexible nature of such MWEs makes it difficult to reliably search for such expressions in text corpora. Standard search engines such as Google do not enable the user to systematically search for different word forms of the same lemma. Search applications for Dutch such as OpenSoNaR (van de Camp et al., 2017; de Does et al., 2017) or Nederlab (Brugman et al., 2016) can do this, but it is difficult to formulate a query allowing different orders and interspersed irrelevant words, and the results of such a query will be unreliable. At best, one can find all instances but one will at the same time find many instances where all these words occur but not the MWE. One should be able to search for a flexible MWE in such a way that its grammatical structure is taken into account. This can be done in a treebank, and MWE-Finder enables this. But MWE-Finder needs as input a MWE in canonical form.

3. Related work

The flexible nature of MWEs makes defining a canonical form (often also called *lemmatisation*)[5] for MWEs challenging. Svensén (2009, 199) notes there are no ready-made solutions in lexicography for representing the different types of variation of idioms. Tiberius and Colman (2023) report on a comparative analysis of lemmatisation practices for MWEs in a number of Dutch general dictionaries and idiom dictionaries. They conclude that "it comes as no surprise that MWEs are not treated consistently at all in the Dutch resources" (p. 2). They focus on dictionaries intended for human users, but for dictionaries intended to be more formal descriptions, the situation is somewhat different. The DuELME database of multiword expressions (Grégoire, 2009, 2010; Odijk, 2013a) contains a quite detailed description of an encoding protocol for MWEs, which has been applied to the 5k MWEs of the DuELME database (Grégoire, 2017). We have made use of this, though the canonical form defined here differs from the one in DuELME.

Geyken (2004) describes a lexical database for German MWEs. The entries contain a 'citation form' (p. 913), but nothing is explicitly said about the form of this citation form. The entries also contain "a POS sequence according to STTS-tagset" with the order of the POS tags following the order in the citation form. From the paper it seems that for verbal MWEs where this is possible the MWE is used with the verb in infinitive form, which is indeed a common use, but what the citation form of other MWE types is remains unclear.

The Modern Greek MWE database IDION (Markantonatou et al., 2019) is addressed to the human user and to NLP. It provides guidelines for the lemma of a MWE, including adopting two 'canonical' orders of phrases in an MWE. This is possible thanks to the relatively free word order of Modern Greek, though the authors deviate from these default orders if a more frequent order exists. We are no experts on

[4] In all examples, we put the MWE components in bold face.

[5] An unfortunate term since most component words of a MWE should not be in lemma form.

the Greek language, but we expect that this description of canonical form is highly underspecified. The PARSEME guidelines[6] speak of a 'prototypical form' and of a 'canonical form' for verbal MWEs (VMWE). These notions are introduced to be able to reconstruct the form to which the PARSEME structural tests must be applied for a given occurrence of a MWE. As for prototypical form, it is defined as follows: "a (candidate) VMWE in its prototypical form (if it exists) is a verbal phrase in active voice whose head verb is in a finite form and whose other lexicalized components depend either on the verb or on another lexicalized component. The VMWE can also contain coordinated verbs." The PARSEME notions of prototypical form and canonical form are different notions than the notion of 'canonical form' used here, since an MWE can have multiple prototypical and canonical forms according to the PARSEME guidelines. Nevertheless, the PARSEME guidelines are very useful to create some order in the occasionally wild MWE variations.

Every MWE lexicon must define how a MWE should be described, but this does not necessarily always require a canonical form (i.e. a natural language expression) for MWEs. So for some MWE databases a canonical form is not relevant, but most of the questions that arise in defining a canonical form will arise in the definition of how a MWE is to be described, as well.

Despite the absence of a precise definition of canonical form for MWEs, several NLP researchers have made MWE lemmatisation tools. Marcińczuk (2017) applies lemmatisation to Polish multi-word common noun phrases and named entities, surely flexible expressions but with limited variation in word order and limited intervention of non-MWE components.

Similarly, Schmitt and Constant (2019) describe a MWE lemmatizer and test it for five languages without explicitly specifying what the lemmatized form of a MWE is (though they do make the occasional remark on it). They seem to assume that the notion of canonical form is clear, which is perhaps justified given the fact that their tool has limitations in that it cannot deal well with changes in word order (p. 143).

4. The DUCAME MWE Resource

The DUtch CAnonicalised Multiword Expressions (DUCAME) lexical resource (Odijk, 2023) is available[7] and consists of a reworked version of the DuELME database (Grégoire, 2009, 2010; Odijk, 2013a), consisting of approximately 5000 entries, and a new list of MWEs (app. 6000 entries) composed on the basis of publicly available sources (Stoett (1923–1925), Onze Taal,[8] VRT website,[9] Lassy-Small treebank (van Noord et al., 2013), own collection). DUCAME is unique in that it has all the MWEs in a canonical form as described in more detail below. The MWEs also have annotations on properties of the MWE components.

DUCAME contains canonical forms for MWEs mainly based on native speaker intuitions on the MWEs by the composer of DUCAME. These native speaker intuitions should be tested against corpus data, because they often underestimate the potential of a MWE. MWE-Finder (Odijk et al., to appear 2024) can be used to search for a particular MWE starting from the canonical form in DUCAME.[10] The results found and the analysis of the search results can then be used to adjust the canonical form for a particular MWE, and to include this adjusted canonical form in a new version of DUCAME.

So far, three versions of DUCAME have been published. The latest (Version 3) contains more than 11k MWEs in canonical form. As stated before, DUCAME has also been integrated in MWE-Finder.

5. A Canonical Form for MWEs

The canonical form for MWEs proposed here:

- defines a single form of the MWE, which can be used as its lexical item. It is important for flexible MWEs to have a canonical form to avoid duplication in a lexicon. The canonical form serves the same function as the lemma that is used as a headword in traditional dictionaries: a single form for a lexical item that can occur in multiple variants.
- implicitly or explicitly specifies how the MWE can vary with regard to inflection, determination and modification of its components

The form of the canonical form for a MWE is determined by a number of guiding principles that we describe in Section 5.1. Principles that determine how to generalize from the properties of the component words in the canonical form are described in Section 5.2. Deviations from these guiding principles can be specified by means of annotations, introduced in Section 5.3. The form that MWE components must take is described in Section 5.4.

[6] https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.2/?page=variants#variants

[7] https://surfdrive.surf.nl/files/index.php/s/2Maw8O0QTPH0oBP

[8] https://onzetaal.nl/schatkamer/lezen/uitdrukkingen and https://onzetaal.nl/zoekresultaten?in=advice&zoek=uitdrukking

[9] https://vrttaal.net/taaladvies-taalkwestie/vaste-uitdrukkingen

[10] Or from a canonical form provided by the user.

There are also rules that determine the nature of the variation of a MWE with regard to inflection, determination and modification in comparison to the canonical form, which are described in Section 5.5. The marking of free arguments and modifiers is described in Section 5.6, the marking of bound pronouns in Section 5.7. A variety of other annotations is discussed in Section 5.8.

5.1. Guiding principles for canonical forms

Canonical forms for MWEs are constructed in accordance with the following guiding principles:

1. An absolute requirement for the canonical form of a MWE is that it is a well-formed expression of the relevant language.

2. In the canonical form, the head of the MWE is in its lemma form if it allows inflectional variants, otherwise in the form that it has in the MWE. Further elaboration of this principle can be found in Section 5.4.

3. Free arguments are represented by indefinite pronouns. Further elaboration of this principle can be found in Section 5.6.

4. Bound arguments are represented by third person singular reflexive or possessive pronouns. See Section 5.7 for elaboration.

5. The head of a MWE is inflectable if in lemma form, non-head components are only inflectable for purely grammatical properties. This is elaborated in Section 5.5.1.

6. The head of a MWE is modifiable and determinable, non-head components of the MWE are in general assumed not to be modifiable or determinable. See Section 5.5.2 for elaboration.

7. Components of the MWE may be annotated with codes to explicitly specify their behavior, if it deviates from these guiding principles or from the grammatical property generalisation principles described in Section 5.2. An overview of the annotations is provided in Section 5.3.

8. The canonical form for a MWE only describes properties that are specific to the MWE. Properties of the MWE that are not specific to the MWE but follow from the grammar of the relevant language are not described explicitly. Section 6 illustrates this principle with several examples.

5.2. Principles for generalising over grammatical properties

In this section we describe the principles behind generalising over grammatical properties from components. This generalisation procedure has been implemented in MWE-Finder.

Words have grammatical properties, usually encoded by means of attribute-value pairs. Some of the properties of the words in the canonical form must be retained, others must be dropped, to obtain a description that can be used to define or, in a search application, to search for allowed variants of the MWE.

The attribute-value pairs to encode properties include attributes for the part of speech,[11] for the lemma of the word, for the actual form of the word in the utterance, and for other grammatical properties, among which we distinguish 3 classes:

Subcategorisation properties: properties used to specify a subcategory of the part of speech, e.g., is a pronoun a demonstrative pronoun or a relative pronoun, is an adposition a preposition or a postposition, is a conjunction a coordinate conjunction or a subordinate conjunction, etc.

Interpretable properties: properties that have an influence on the meaning of the utterance, e.g., is a noun singular or plural, what is the mood of the verb, what is the tense of a finite verb, etc. We count oblique case here as well.

Purely grammatical properties: e.g., the person and number of a finite verb, the inflectional form of an adjective, the non-oblique case of a noun or pronoun, etc.

The following principles apply:

- The properties for part of speech, grammatical relation, and lemma, as well as subcategorisation properties of a component are always fixed.
- Components that are in the lemma form and are either the head of the MWE or a non-head component marked with + can have variation in their interpretable and purely grammatical properties.
- Other components of the MWE must keep the lemma of the word, its part of speech, any relevant subcategorisation properties and interpretable properties fixed, but they allow variation for purely grammatical properties. The

[11] In some implemented grammatical frameworks words also have a property for the grammatical relation the word has in the structure. But this is not a property of the word but rather of the relation that the word has to some other word (or to the phrase that it belongs to).

values of these properties are generally not freely selectable but determined by the grammar of the language.

For these other components, variation in purely grammatical properties must be allowed. For example, in MWEs consisting of an adjective and a noun, the adjective must agree with the head noun, and thus it gets its normal inflectional variants in plural and in definite noun phrases, as illustrated in (2):[12]

(2) a. een **vrolijk**-(*e) **Fransje**
a gay-E Frans.DIM
'a gay spark'

b. **vrolijk**-*(e) **Fransjes**
gay-E Frans.DIM.PL
'gay sparks'

c. dit **vrolijk**-*(e) **Fransje**
this gay-E Frans.DIM
'this gay spark'

5.3. Annotations

Exceptions to the guiding and generalisation principles can be specified by means of annotations. The annotations allowed are given in Table 1. They will be explained in more detail in the text below.

5.4. The form of MWE components in the canonical form

For single words the canonical form is called the *lemma*, i.e. a specific form of an inflectional paradigm as found as headword in traditional dictionaries. One can adopt this usage for the heads of MWEs as well, and that works fine for many MWEs. However, it does not always work for MWEs with a verb as its head. For this reason we treat MWEs in two separate sections, one for the canonical form of non-verbal MWEs (Section 5.4.1), and one for the canonical form of verbal MWEs (Section 5.4.2). For non-head components of MWEs, a specific proposal is made in Section 5.4.3.

5.4.1. Nonverbal MWEs

For many MWEs the canonical form is equal to the only form that exists:

(3) **als de wiedeweerga**
like the wiedeweerga
'like greased lightning'

For the head of an MWE the lemma is used unless the head can only appear in a different form. In those cases the actual form is used.
Example where the head is a lemma in the canonical form:

(4) **bijvoeglijk naamwoord**
adjectival nominal
'adjective'

Example where the head is not the lemma:

(5) **Eva's *dochteren***
Eve's daughters
'Eve's daughters' (i.e. 'women')

We assume that comparatives, superlatives and diminutives are derived by derivation (not inflection).[13] They have their own lemma, which differs from the lemma of the word they have been derived from. So, in the following example, the MWE head is in lemma form:

(6) **een illusie *armer***
an illusion poorer
'robbed of an illusion'

5.4.2. Verbal MWEs

In Dutch, the lemma of a verb is identical to the infinitive,[14] but several problems arise when one tries to use the infinitive as the lemma for a verbal MWE. First, no overt subjects can appear with an infinitive, so a MWE with an overt subject and an infinitive is an ill-formed expression:

(7) * **De laatste loodjes het zwaarst wegen**.
the last lead.DIM.PL the heaviest weigh
'The tail end is the most difficult.'

Furthermore, though the subject must be absent, it is present implicitly and interpreted as an animate actor. If the subject of a MWE is not animate, using the MWE with an infinitival head as the canonical form gives infelicitous results:

(8) ? iemand **de keel uithangen**
someone the throat out.hang
'for something to bore someone'

In order to avoid these problems and at the same time have a canonical form with an infinitive, the canonical forms are all finite sentences with a finite auxiliary verb as its main head. For Dutch, a form

[12] The notation *(...) means that leaving out the parts between the round brackets yields ill-formedness, the notation (*...) means that including the part between the brackets leads to ill-formedness. E stands for the adjectival *e*-suffix. *Frans* is a common Dutch name. DIM stands for diminutive, PL for plural.

[13] This is not generally accepted, and the grammar behind the parser that is used in GrETEL (Alpino (van Noord, 2006)) treats these phenomena as inflection.

[14] This may be different in other languages. In such languages different conventions will have to be assumed.

notation	interpretation
**word*	*word* is modifiable/determinable
+*word*	*word* is inflectable
=*word*	*word* must occur in the MWE as given
!*word*	*word* is not modifiable/determinable
dd:[*word*]	*word* must be a definite determiner
<*text*>	*text* is interpreted as a freely replaceable argument
0*word*	*word* is not part of the MWE

Table 1: Notational devices for annotating a canonical form. The code + can also be combined with * or *!* (in any order).

of the future tense auxiliary verb *zullen* 'will' has been selected for this purpose, as in (9).[15] These are all well-formed sentences that can in principle be parsed by a parser.

(9) a. **De laatste loodjes** zullen **het** the last lead.DIM.PL will the **zwaarst wegen.** heaviest weigh
'The tail end will be the most difficult.'

b. Iets zal iemand **de keel** something will someone the throat **uithangen**. out.hang
'Something will bore someone.'

The forms of the auxiliary *zullen* are completely ignored by MWE-Finder when generalising from the canonical form to a query to search for the relevant MWE.

5.4.3. Non-head components

For non-head components of the MWE, if only one form occurs, that form is used:

(10) a. ***brave*** **Hendrik** good Henry
'good soul'

b. **gouden appels op zilveren *schalen*** gold apples on silver plates
'gold apples on silver plates'

If non-head components can occur in multiple forms, the general rule is that a form should be selected that yields a well-formed expression. In Dutch, e.g., non-head adjectival modifiers to nouns can sometimes occur in multiple forms. Use the form as used in a singular indefinite noun phrase: ***vrolijk*** *Fransje*, not *(dit)* ***vrolijke*** *Fransje* or ***vrolijke*** *Fransjes*, because this is the only form that yields a well-formed expression in combination with the lemma form of the head noun.

[15] The requirement to mark this auxiliary with a preceding *0* is under consideration.

5.5. Inflection and modification of components

In this section we describe what is assumed as a default for head and non-head components with regard to inflection and modification. There are many exceptions to these defaults, and these can be indicated by means of annotations.

5.5.1. Inflection of components

It is assumed that the head of the MWE can occur in all inflected forms if it equals the lemma form. Exceptions to this can be marked by the symbol = in front of the word. With such a marking only the listed form is allowed.
Since diminutives, comparatives and superlatives are assumed to be created by derivation from the base form, the diminutive, comparative and superlative suffixes are part of the lemma and the features encoding these forms must be retained, but other inflectional affixes can be added (e.g., for plural, the *e*-suffix on adjectives, etc.).
The non-head components can only occur in the form listed, except for variations based on purely grammatical features, or when they are preceded by +: in that case all inflected forms are allowed (to the extent that they are grammatically possible in the context).

5.5.2. Modification of components

It is assumed that the head of a MWE can occur with modifiers and determiners that are not components of the MWE. Exceptions can be marked by the symbol ! in front of the head word.
Concerning non-head components, the canonical forms must be interpreted as not allowing modifiers or determiners that are not components of the MWE. Exceptions to this can be marked by the symbol * in front of the word.

(11) Iemand zal ***terrein verliezen**. someone will territory lose
'Someone will lose ground.'

5.6. Free arguments and modifiers

Arguments of the MWE that can be freely replaced by arbitrary phrases are represented by the in-

definite pronouns *iemand* 'someone', *iets* 'something', and *ergens* 'somewhere', where this is possible. One can also use combinations such as *iemand*|*iets* or *iets*|*iemand*, which are to be interpreted as allowing either but most likely the first alternative. The presence of the arguments is required. We follow here the recommendation of (Svensén, 2009) that MWEs must be presented in their full form and in their usual constructions, i.e. the syntactic valency of the MWE must be shown. There is no notation for optional arguments (except for comitative arguments), so optional arguments require two separate canonical forms.
If indefinite pronouns must occur as such in the MWE (i.e. cannot be freely replaced), one can have them preceded by the annotation =, as in (12).

(12) Iemand zal **voor** **=iets** **tussen** iets **zitten**.
someone will for something between something sit
'Someone will be a factor in something.'

Many PP's can occur either before the verb (cluster) or after it in Dutch. In the canonical form the PP is always to the left of the verb. Furthermore, the PP can often be in multiple positions when it is to the left of the verb. In the canonical form it must be as close to the verb as possible, again to obtain a unique canonical form:

(13) a. Iemand zal ***behoefte aan** iets **hebben**. (OK)
b. Iemand zal ***behoefte hebben aan** iets. (NOT OK)
c. Iemand zal **aan** iets ***behoefte hebben**. (NOT OK)

5.6.1. Comitative arguments

A comitative argument is introduced by the preposition *met* and is optional, unless the co-argument subject or object is singular and does not denote a group or aggregate.
The co-argument is indicated by *iemand*, as usual. The preposition *met* is marked with com:[].

(14) Iemand zal **een** **blik** com:[met] iemand **wisselen**.
someone will a look with someone exchange
'Someone will exchange looks with someone.'

Currently, there is no check on the semantics in MWE-Finder, so using the *com:[]* annotation will simply allow the MWE to occur with the *met*-phrase or without.

5.6.2. Possessive free arguments

Possessive arguments are indicated with *iemands*:

(15) Iets zal **door** iemands **hoofd** **malen**.
something will through someone's head grind
'Something will keep running through someone's head.'

Such possessive arguments are interpreted as allowing a number of variants, e.g. for *iemands hart*: <possessive NP> *hart*, e.g. *mijn tantes hart*, 'my aunt's heart'; *het hart van* <NP>, e.g. *het hart van de buurman*, 'the heart of the neighbour'; <NP> *z'n/d'r hart*, e.g. *haar vriend z'n hart*, lit. 'her friend his heart'; and <possessive pronoun> *hart*, e.g. *mijn hart*, 'my heart'.

5.6.3. Free arguments in PPs

Free arguments to an adposition are indicated with the usual indefinite pronouns.
A prepositional phrase with *iets* as complement to an preposition can occur in two variants. For example, *over iets* lit. *about something* and *ergens over* lit. 'somewhere about' are both well-formed Dutch expressions meaning 'about something'. In canonical forms inanimate free arguments are always indicated with *iets*, never with *ergens*:

(16) a. Iemand zal ***behoefte** **aan** iets **hebben**. (OK)
someone will need to something have
'Someone will have a need for something.'
b. Iemand zal ergens ***behoefte** **aan** **hebben**. (NOT OK)
Someone will somewhere need to have

One form must be selected to have a unique canonical form, and we opted for P + *iets* because *ergens* 'somewhere' is used for free locative arguments and modifiers.
Prepositional complements to verbs with *iets* also give rise to variants with R-pronouns separated from the adposition (e.g., *ergens aan*), and R-pronouns written with the adposition as a single word (e.g., *hieraan*, etc., the so-called 'pronominal adverbs').

(17) a. Iemand zal daar geen **behoefte** **aan** **hebben**.
someone will there no need to have
'Someone will have no need for that.'

b. Iemand zal daar**aan** geen **behoefte** **hebben**.
someone will there.to no need have
'Someone will have no need for that.'

It also gives rise to variants where additionally a sentential complement is present:

(18) Iemand zal er **behoefte aan hebben** om te vertrekken.
someone will there need to have for to leave
'Someone will have the need to leave.'

5.6.4. Free arguments other than NPs

For many phrase types other than NPs there are no pronouns at all, e.g. for adjectival, adverbial and clausal phrases.[16] The missing pronouns are covered by a special annotation in which an arbitrary phrase surrounded by angled brackets (<...>) is interpreted as a freely replaceable argument, as in (19). This is especially relevant for adjectives and adverbs acting as modifiers, for directional PPs as modifier or locational-directional complement, and for certain determiners. The angled brackets are interpreted as an arbitrary phrase with the particular grammatical relation but no further restrictions are imposed.

(19) Iemand zal <makkelijk> **in de omgang zijn**.
someone will easy in the interaction be
'Someone will be easy-going'

A special case is formed by phrases that can occur only as a predicate. Here the (copula) verb is an open slot:

(20) Iemand zal **de dupe** <zijn>.
someone will the victim <be>
'Someone will be the victim.'

If an argument must be plural or collective, then the use of *iemand* leads to an unnatural sentence. For such cases one can use the plural pronoun *zij* between angled brackets
We are working on restrictions on what kind of expressions can be used with angled brackets to ensure that the canonical form is indeed unique.

(21) <Zij> zullen **uit elkaar gaan**.
they will out.of each.other go
'They will break up their marriage.'

[16] One can sometimes use pronouns for noun phrases to refer to these but there are no pronouns that can actually replace them.

5.7. Bound Pronouns

Bound pronouns such as reflexives and possessive pronouns are represented by the third person singular forms (*zich*, *zichzelf*, *zijn*). There is (currently) no convention or annotation to specify the antecedent of such bound anaphors.

(22) a. Iemand zal **de schepen achter zich verbranden**.
someone will the ships behind SELF burn
'Someone will burn his boats.'
b. Iemand zal **zijn gram halen**.
someone will his anger fetch
'Someone will get square.'

These are interpreted as allowing all variants of *zich* (*me*, *je*, etc.), *zichzelf* (*mijzelf*, *mezelf*, *jezelf*, etc.), and *zijn* (*mijn*, *jouw*, etc.). If such forms do not vary, one can precede them by the annotation =, as in the expressions in (23).

(23) a. **op =zich**
on SELF
'in itself'
b. **op =zijn elfendertigst**
on his eleven.and.thirtiest
'at a snail's pace'

5.8. Other annotations

5.8.1. Definite determiner variation

The code *dd:[...]* indicates that the word between the square brackets can be replaced by an arbitrary definite article or demonstrative pronoun. i.e. *het, dit, dat*; or *de, deze, die*.[17]

(24) **dd:[dat] oude liedje**
that old song.DIM
'the same old story'

5.8.2. Zero elements

One sometimes has to include a word in a canonical form to create a natural utterance even if this word does not belong to the MWE. Such words can be preceded by the code *0*. This very often occurs in MWEs that have an indefinite inanimate subject, which prefer the presence of *er*, as in (25a), and for determiners required on count nouns, as in (25b).

(25) a. 0Er zal iets **op het spel staan**.
there will something on the game stand
'Something will be at stake.'

[17] Whether *de* or *het* can appear is not a property of the MWE but of the grammar of Dutch.

b. Iemand zal 0dat *+**varkentje**
someone will that pig.DIM
wassen.
wash
'Someone will solve that problem.'

5.8.3. Negative Polarity Licensors

Negative Polarity MWEs require a Negative Polarity Licensor such as a negative adverb (*niet* 'not'), determiner (*geen* 'no') or pronoun (e.g. *niemand* 'nobody'). In (26a) the negative adverb *niet* cannot be absent, but it is arguably not part of the MWE, as shown by (26b) in which the negative pronoun *niemand* in the main clause is the licensing element for the negative polarity MWE in the subordinate clause.

(26) a. Die **vlieger** zal *(niet) **opgaan**.
that kite will not up.go
'That won't wash.'

b. Niemand denkt dat die **vlieger**
nobody thinks that that kite
opgaat.
up.goes
'Nobody thinks that that will wash.'

Negative Polarity Licensors are marked with the symbol *0* in front of the word:

(27) a. Die **vlieger** zal 0niet **opgaan**.
that kite will not up.go
'That won't wash.'

b. Iemand zal **er** 0geen
someone wil there no
doekjes **om** **winden**.
tissue.DIM.PL around wind
'Someone will not beat about the bush.'

6. Properties not specific to the MWE

There are many properties that are true of a MWE but that are not specific to this MWE but follow from the grammar of the language. For example, in the MWE *vrolijk Fransje* the adjective *vrolijk* precedes the noun *Fransje*, but this is not a property specific to this MWE but rather a property of the grammar of Dutch, which requires attributive adjectives to precede the noun they modify. Left-right order is generally not a property specific to a MWE, with the exception of the order of conjuncts in a coordinate structure: here the grammar of Dutch imposes no order but a specific MWE may do so, as in (28):[18]

(28) a. **dames en heren**
ladies and gentlemen
'ladies and gentlemen'

b. # heren en dames
gentlemen and ladies
'gentlemen and ladies'

Dutch has words that in some cases must be used as a preposition (preceding its complement) and in other cases as a postposition (following its complement), but even this does not require conditions on order since we assume that this distinction is marked by a grammatical feature (as indeed it is in the grammar used in MWE-Finder). Thus, even without restrictions on left-right order, (29) will correctly not be identified as containing the MWE *op de klippen lopen* 'to fail', though (30) will:

(29) Dat zal de klippen op lopen.
that will the cliffs on walk
'That will walk onto the cliffs.' (Not: 'That will fail.')

(30) Dat zal **op de klippen lopen**.
that will on the cliffs walk
'That will walk on the cliffs.' And: 'That will fail.'

Another example of properties that are generally not properties of the MWE but of the grammar of Dutch concerns the value of purely grammatical properties (e.g., person and number on verbs, inflection on attributive adjectives). See (2) in Section 5.2 for illustration.

7. Concluding Remarks

This paper proposed design principles and guidelines for a canonical form for (flexible) Multiword Expressions (MWEs) and introduced a lexical resource, DUCAME, which contains more than 11k Dutch multiword expressions in the canonical form as defined in this paper. It elaborated the principles and guidelines in detail for the Dutch language. The DUCAME resource is used in MWE-Finder as a basis for generating queries to search for occurrences of an MWE in large text corpora.
This version of the canonical form lacks many desired features. These involve features for collocations, support verb constructions, the meaningfulness of certain components, and other aspects. We hope to extend the design principles and guidelines for these aspects, together with a new version of DUCAME, in future work.

Acknowledgements

We thank Tijmen Baarda, Ben Bonfil, and Sheean Spoel and the anonymous reviewers of this paper for their comments on an earlier version of this paper, which has led to many improvements. Of course, all errors are ours.

[18] The symbol # means that the utterances is well-formed but has no interpretation as MWE.

8. Bibliographical References

Hennie Brugman, Martin Reynaert, Nicoline van der Sijs, René van Stipriaan, Erik Tjong Kim Sang, and Antal van den Bosch. 2016. Nederlab: Towards a single portal and research environment for diachronic Dutch text corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Jesse de Does, Jan Niestadt, and Katrien Depuydt. 2017. Creating research environments with BlackLab. In Jan Odijk and Arjan van Hessen, editors, *CLARIN in the Low Countries*, chapter 19, pages 245–257. Ubiquity, London, UK. DOI: http://dx.doi.org/10.5334/bbi.20. License: CC-BY 4.0.

Alexander Geyken. 2004. Bootstrapping a database of German multi-word expressions. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Nicole Grégoire. 2009. *Untangling Multiword Expressions: A study on the representation and variation of Dutch multiword expressions*. Phd, Utrecht University, Utrecht. LOT Publication.

Nicole Grégoire. 2010. DuELME: A Dutch electronic lexicon of multiword expressions. *Journal of Language Resources and Evaluation*, 44(1/2):23–40.

Nicole Grégoire. 2017. MWE lexicon for Dutch: Encoding protocol. STEVIN IRME report, Utrecht University, Utrecht. Part of the DuELME documentation.

Michał Marcińczuk. 2017. Lemmatization of multiword common noun phrases and named entities in Polish. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 483–491, Varna, Bulgaria. INCOMA Ltd.

Stella Markantonatou, Panagiotis Minos, George Zakis, Vassiliki Moutzouri, and Maria Chantou. 2019. IDION: A database for Modern Greek multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 130–134, Florence, Italy. Association for Computational Linguistics.

Jan Odijk. 2013a. DUELME: Dutch electronic lexicon of multiword expressions. In G. Francopoulo, editor, *LMF - Lexical Markup Framework*, pages 133–144. ISTE / Wiley, London, UK / Hoboken, US.

Jan Odijk. 2013b. Identification and lexical representation of multiword expressions. In P. Spyns and J.E.J.M Odijk, editors, *Essential Speech and Language Technology for Dutch. Results by the STEVIN-programme*, Theory and Applications of Natural Language Processing, pages 201–217. Springer, Berlin/Heidelberg.

Jan Odijk, Martin Kroon, Tijmen Baarda, Ben Bonfil, and Sheean Spoel. to appear 2024. MWEfinder: Querying for multiword expressions in large Dutch text corpora. In Voula Giouli and Verginica Barbu Mititelu, editors, *Multiword expressions in lexical resources. Linguistic, Lexicographic and Computational perspectives*, Phraseology and Multiword Expressions. Language Science Press, Berlin.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2001. Multiword expressions: A pain in the neck for NLP. *LinGO Working Paper*, 2001-03.

Marine Schmitt and Mathieu Constant. 2019. Neural lemmatization of multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 142–148, Florence, Italy. Association for Computational Linguistics.

Frederik August Stoett. 1923–1925. *Nederlandsche spreekwoorden, spreekwijzen, uitdrukkingen en gezegden*, 4th edition. W.J. Thieme & Cie, Zutphen.

Bo Svensén. 2009. *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making.* Cambridge University Press.

Carole Tiberius and Lut Colman. 2023. Lemmatisation of MWEs in Dutch resources. In *Proceedings of the First General UniDive Meeting*, Paris-Saclay, France. UniDive.

Matje van de Camp, Martin Reynaert, and Nelleke Oostdijk. 2017. WhiteLab 2.0: A web interface for corpus exploitation. In Jan Odijk and Arjan van Hessen, editors, *CLARIN in the Low Countries*, chapter 19, pages 231–243. Ubiquity, London, UK. DOI: http://dx.doi.org/10.5334/bbi.19. License: CC-BY 4.0.

Gertjan van Noord. 2006. At last parsing is now operational. In *TALN06 Verbum Ex Machina. Actes de la 13e conference sur le traitement automatique des langues naturelles*, pages 20–42, Leuven, Belgium.

Gertjan van Noord, Gosse Bouma, Frank Van Eynde, Daniël de Kok, Jelmer van der Linde, Ineke Schuurman, Erik Tjong Kim Sang, and Vincent Vandeghinste. 2013. Large scale syntactic annotation of written Dutch: Lassy. In Peter Spyns and Jan Odijk, editors, *Essential Speech and Language Technology for Dutch*, Theory and Applications of Natural Language Processing, pages 147–164. Springer Berlin Heidelberg.

9. Language Resource References

Jan Odijk. 2023. *DUCAME: DUtch CAnonicalised Multiword Expression Database*. Utrecht University. Utrecht University, 3.0.

Jan Odijk, Martin Kroon, Tijmen Baarda, Ben Bonfil, and Sheean Spoel. 2023. *MWE-Finder. Application to identify Dutch MWEs*. Utrecht University. Utrecht University. Github: https://github.com/UUDigitalHumanitieslab/mwe-query and https://github.com/UUDigitalHumanitieslab/gretel.